Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2023;15(2):

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2023-96566

Hasta Bazlı Heterojenitenin Giderilmesi: Kümeleme Analizi Yaklaşımı

Elimination of Patient-Based Heterogeneity: Cluster Analysis Approach

Batuhan BAKIRARAR[a]

[a]Ankara Üniversitesi Tıp Fakültesi, Biyoistatistik ABD, Ankara, Türkiye

ÖZET Amaç: Çalışmanın amacı, uygun kümeleme yöntemleri ve küme sayıları kullanarak verileri uygun alt gruplara bölmek, ilgilenilen bağımlı değişken için analizleri bu alt gruplarda yaparak sınıflama performansındaki değişimleri incelemektir. **Gereç ve Yöntemler:** Çalışmada kullanılan veri seti Kaggle veri tabanından alınmıştır. Veri seti, 15 bağımsız ve 1 bağımlı değişken olmak üzere toplam 16 değişkenden oluşmakta olup, 948 hasta verisi içermektedir. Analizler R (ver. 4.1.2) programla dili kullanılarak yapılmış ve bu programla dili içindeki chisquare, wilcox.test, cluster, RWeka ve e1071 paketlerinden faydalanılmıştır. Kümeleme yöntemi olarak K-ortalama, sınıflandırma yöntemleri olarak ise lojistik regresyon, çok katmanlı algılayıcı, karar tablosu, karar destek makinesi, rastgele orman, J48, IBk, torbalama ve oylanmış algılayıcı kullanılmıştır. Veri seti 10 kat çapraz doğrulama kullanılarak test edilmiş ve tüm analizler 1.000 kez tekrarlanarak sonuçlar verilmiştir. Performans kriteri olarak doğru sınıflama oranı, F-ölçütü, Matthews korelasyon katsayısı, ROC eğrisi altında kalan alan ve precision-recall alanı kullanılmıştır. **Bulgular:** Modeller ile Tip 2 diyabet olacak ve olmayacak hastalar için tahmin sonuçlarına bakıldığında, birinci kümede Tip 2 diyabet görülecek tahmini yapılan hastalara ait performansta genel modele göre artış gözlenmiştir. İkinci kümede Tip 2 diyabet görülmeyecek tahmini yapılan hastalara ait performansta da genel modele göre artış gözlenmiştir. **Sonuç:** Çalışmada yapılan kümeleme analizi sonrasında hastalara ait spesifik özelliklerin daha iyi belirlendiği ve küme bazlı sınıflandırma ile de oluşturulan modellerinin daha iyi sınıflama performansına sahip olduğu gözlenmiştir.

Anahtar kelimeler: Kümeleme; K-ortalama; sınıflandırma; makine öğrenmesi; heterojenite

ABSTRACT Objective: The aim of the study is to divide the data into appropriate subgroups using appropriate clustering methods and cluster numbers, and to analyze the changes in classification performance by performing the analyzes for the dependent variable of interest in these subgroups. **Material and Methods:** The data set used in the study was taken from the Kaggle database. The data set consists of 16 variables, 15 independent and 1 dependent variable, and includes 948 patient data. For the analysis, the R (ver. 4.1.2) programming language was used and the chisquare, wilcox.test, cluster, RWeka and e1071 packages in this program were used. K-means was used as the clustering method, and logistic regression, multilayer perceptron, decision table, decision support machine, random forest, J48, IBk, bagging and voted perceptron were used as classification methods. The dataset was tested using 10-fold cross validation option and all analyzes were repeated 1,000 time. Accuracy, F-measure, Matthews correlation coefficient, area under ROC curve and precision-recall area were used as performance criteria. **Results:** Considering the estimation results for patients with and without Type 2 diabetes with the models, an increase was observed in the performance of patients who were predicted to have Type 2 diabetes in the first cluster compared to the general model. An increase was observed in the performance of patients who were predicted not to have Type 2 diabetes in the second cluster compared to the general model. **Conclusion:** After clustering, it was observed that the specific characteristics of the patients were determined better and the models created with the cluster-based classification had better performance.

Keywords: Clustering; K-means; classification; machine learning; heterogeneity

Correspondence: Batuhan BAKIRARAR
Ankara Üniversitesi Tıp Fakültesi, Biyoistatistik ABD, Ankara, Türkiye
E-mail: batuhan_bakirarar@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 07 Mar 2023 ***Received in revised form:*** 29 May 2023 ***Accepted:*** 30 May 2023 ***Available online:*** 02 Jun 2023

Veri heterojenliği, çok farklı dağılımlara sahip verileri ifade eder. Tıp merkezlerindeki hasta popülasyonları, çevre, prosedürler ve tedavi protokolleri farklılık gösterdiğinden, heterojenlik genellikle sağlık hizmeti verilerinde ortaya çıkar.[1] Bu heterojenlik, nüfusa ve çevresel etkenlere bağlı olmak üzere 2 grupta incelenebilir. Nüfusa bağlı heterojenliğe, hastaların genetik mirası, yaşı, cinsiyeti, hastalık şiddeti ve bireysel sosyoekonomik durumu neden olabilir. Örneğin 65 yaşından büyük travmatik beyin hasarı olan hastalarda 6 aylık ölüm oranı yaklaşık %72 iken, 35 yaşından genç hastalarda bu oran %21'dir. Çevresel heterojenlik, farklı hastanelerin fiziksel konumlarından kaynaklanmaktadır. Dünyanın çeşitli bölgelerinin hava durumu, hava kalitesi, sıcaklığı ve mevsimleri veri kümelerinde heterojenliğe neden olabilir. Örneğin hastaların maruz kaldığı güneş miktarı melanom görülme riskini artırmaktadır.[2,3]

Hastaların heterojenliği sorunuyla başa çıkmak için ise çözüm yolları mevcuttur. Heterojenliği ortadan kaldırmak için analiz aşamasında genellikle 2 yöntem kullanılır: ortak değişken düzeltmesi ve alt grup analizi. Ortak değişken düzeltmesi, düzeltilmemiş tahminlerin aksine daha spesifik ve daha doğru sonuçlar elde edilmesine yardımcı olur. Tedavi etkisinin düzeltilmemiş bir tahmini, belirli bir hastalığı olan "ortalama" bir hastayla ilgili bir yorum sağlarken, bir öngörücü (örneğin yaş) için düzeltme yapmak ise "belirli bir yaştaki bir hasta" için kestirim yapılmasına olanak sağlar. Ortak değişken düzeltmesi ayrıca dengesizliği azaltır ve önemli bir tedavi etkisini saptamak için istatistiksel gücü artırır. Böylece potansiyel olarak örneklem büyüklüğü gereksinimlerini azaltır.[4] Örneğin GUSTO-I çalışmasına dâhil edilen akut miyokard infarktüslü hastalarda tedavi etkisine 17 ortak değişken ile düzeltme yapılarak bakılması, gerekli örneklem büyüklüğünü %15 azaltmıştır.[5] Travmatik beyin hasarı verileri baz alınarak yapılan bir simülasyon çalışmasında, yaş ve Glasgow Koma Skalası motor skoru için yapılan düzeltme, gerekli örneklem büyüklüğünü %30 azaltmıştır.[6]

Alt grup analizleri, genellikle aralarında karşılaştırmalar yapmak için tüm katılımcı verilerini alt gruplara ayırmayı içerir. Alt grup analizleri, katılımcıların alt kümeleri (erkekler ve kadınlar gibi) veya çalışmaların alt kümeleri (farklı coğrafi konumlar gibi) için yapılabilir. Alt grup analizleri, heterojen sonuçları araştırmak veya belirli hasta grupları, müdahale türleri veya çalışma türleri hakkında belirli soruları yanıtlamak için yapılabilir.[4]

Her ne kadar aynı evrenden örneklem toplansa, veri yapısının homojen olmasına dikkat edilse bile hasta yapısındaki heterojenlik nedeniyle bu hastaları yaş, cinsiyet, ek hastalık vb. risk faktörlerine göre alt gruplara (kümelere) ayırmadan analiz etmek elde edilen sonuçlarda yanlışlığa yol açar.[2] Son yıllarda yapılan çalışmalarda, hastaların ortak özelliklerine göre kümelere ayrılması, analizlerin bu alt gruplarda ayrı ayrı yapılması ve sonuçların daha sonra gerekirse birleştirilmesinin önemi vurgulanmıştır. Bu sayede hasta yapısındaki heterojenliğin giderilebileceği ve sonuçların daha doğru olacağı, popülasyonu temsil eden daha iyi modellemeler elde edilebileceği vurgulanmıştır.[4,7]

Çalışmanın amacı, uygun kümeleme yöntemleri ve küme sayıları kullanarak verileri uygun alt gruplara bölmek, ilgilenilen bağımlı değişken için analizleri bu alt gruplarda yaparak sınıflama performansındaki değişimleri incelemektir.

GEREÇ VE YÖNTEMLER

Çalışma, metodolojik bir araştırma olarak planlanmıştır.

VERİ SETİ

Orijinal veri seti, Kaggle veri tabanından alınmış olup, Tip 2 diyabet görülen ve görülmeyen toplam 952 hastayı içermektedir.[8] Veri seti, 15 bağımsız (yaş, cinsiyet, ailede diyabet öyküsü, yüksek kan basıncı teşhisi, fiziksel aktivite, beden kitle indeksi, sigara tüketimi, alkol tüketimi, günlük uyku süresi, günlük derin uyku süresi, düzenli ilaç alımı, abur cubur tüketimi, stres, kan basıncı seviyesi, idrara

çıkma sıklığı) ve 1 bağımlı (Tip 2 diyabet) değişken olmak üzere toplam 16 değişkenden oluşmaktadır. Veri setindeki eksik veri içeren 4 hasta çıkarıldıktan sonra çalışma 948 hastalık örneklem ile yapılmıştır.

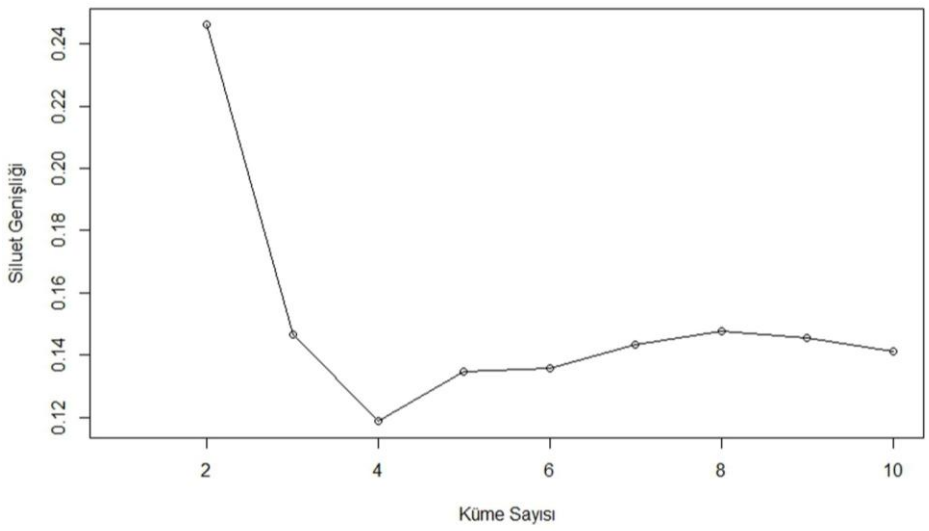
İSTATİSTİKSEL ANALİZ

Tanımlayıcı olarak nicel değişkenler için ortalama±standart sapma ve ortanca (minimum-maksimum), nitel değişkenler için ise hasta sayısı (yüzde) kullanılmıştır. Nicel değişken bakımından iki kategoriye sahip nitel değişkenin kategorileri arasında fark olup olmadığına, normal dağılım varsayımları sağlanmadığı için Mann-Whitney U testi kullanılarak bakılmıştır. İki nitel değişken arasındaki ilişki incelenmek istendiğinde ise Pearson ki-kare testi kullanılmıştır. Veri setini kümelere bölmek için K-ortalama kümeleme algoritması kullanılmış ve bölünecek ideal küme sayısı 2 olarak bulunmuştur. Değişkenlerin önemini belirlemek için "gain ratio attribute evaluation testi" kullanılmıştır. Makine öğrenmesi sınıflama yöntemlerinden lojistik regresyon, çok katmanlı algılayıcı (gizli katman 3, öğrenme oranı: 0,3 ve momentum: 0,2), karar tablosu (decision table), karar destek makinesi (seed: 1, tolerans: 0,001), rastgele orman (ağaç sayısı 100 ve seed: 2), J48, IBk, torbalama (bagging) (seed: 1) ve oylanmış algılayıcı (voted perceptron) (seed: 1, kuvvet: 1) kullanılmıştır. Veri seti 10 kat çapraz doğrulama kullanılarak test edilmiş ve tüm analizler 1.000 kez tekrarlanarak sonuçlar verilmiştir. Performans kriteri olarak doğru sınıflama oranı (DSO), F-ölçütü, Matthews korelasyon katsayısı, ROC eğrisi altında kalan alan ve precision-recall alanı kullanılmıştır. Tüm analizler R (ver. 4.1.2) programla dili kullanılarak yapılmış ve bu programla dili içindeki chisquare, wilcox.test, cluster, RWeka ve e1071 paketlerinden faydalanılmıştır.

BULGULAR

Tablo 1'de genel veri seti için Tip 2 diyabet varlığı ile değişkenlere ait karşılaştırmalar yer almaktadır. Karşılaştırmalara bakıldığında cinsiyet (p=0,244), sigara tüketimi (p=0,786), günlük uyku süresi (p=0,177), günlük derin uyku süresi (p=0,725) ve abur cubur tüketimi (p=0,131) dışındaki tüm değişkenler için diyabet varlığı bakımından anlamlı fark bulundu (p<0,05).

Çalışmadaki amacımız doğrultusunda veri seti ideal küme sayısı olarak belirlenen 2 kümeye bölündü ve her bir küme için tüm analizler ayrı ayrı uygulandı (Şekil 1). Birinci kümede 523, ikinci kümede ise 425 hasta bulunmaktaydı. Kan basıncı seviyesi ve yüksek kan basıncı teşhisi ilişkili değişkenler olduğu için analizlere daha çok bilgi sağladığı düşünülen kan basıncı seviyesi dâhil edildi.



ŞEKİL 1: İdeal küme sayısı için analiz sonuçları.

TABLO 1: Diyabet için değişkenlere ait karşılaştırmalar.

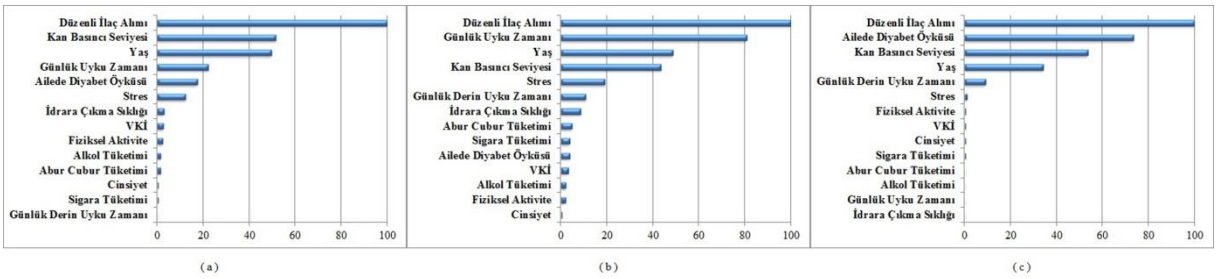
Değişkenler		Tip 2 Diyabet			p değeri
		Yok (n=683)	Var (n=265)	Toplam (n=948)	
Yaş, n (%)	<40	453 (93,4)	32 (6,6)	485 (51,2)	**<0,001**[a]
	40-49	114 (69,5)	50 (30,5)	164 (17,3)	
	50-59	84 (54,2)	71 (45,8)	155 (16,4)	
	≥60	32 (22,2)	112 (77,8)	144 (15,2)	
Cinsiyet, n (%)	Kadın	258 (69,9)	111 (30,1)	369 (38,9)	0,244[a]
	Erkek	425 (73,4)	154 (26,6)	579 (61,1)	
Ailede diyabet öyküsü, n (%)	Yok	412 (83,1)	84 (16,9)	496 (52,3)	**<0,001**[a]
	Var	271 (60,0)	181 (40,0)	452 (47,7)	
Yüksek kan basıncı teşhisi, n (%)	Yok	588 (81,4)	134 (18,6)	722 (76,2)	**<0,001**[a]
	Var	95 (42,0)	131 (58,0)	226 (23,8)	
Fiziksel aktivite, n (%)	Yok	77 (58,8)	54 (41,2)	131 (13,8)	**0,001**[a]
	<0,5 saat	253 (75,3)	83 (24,7)	336 (35,4)	
	0,5 saat-1 saat	205 (76,2)	64 (23,8)	269 (28,4)	
	≥1 saat	148 (69,8)	64 (30,2)	212 (22,4)	
BKİ, n (%)	Zayıf	48 (85,7)	8 (14,3)	56 (5,9)	**0,002**[a]
	Normal	315 (75,5)	102 (24,5)	417 (44,0)	
	Fazla kilolu	184 (70,2)	78 (29,8)	262 (27,6)	
	Obez	136 (63,8)	77 (36,2)	213 (22,5)	
Sigara tüketimi, n (%)	Yok	604 (71,9)	236 (28,1)	840 (88,6)	0,786[a]
	Var	79 (73,1)	29 (26,9)	108 (11,4)	
Alkol tüketimi, n (%)	Yok	556 (73,5)	200 (26,5)	756 (79,7)	**0,041**[a]
	Var	127 (66,1)	65 (33,9)	192 (20,3)	
Günlük uyku süresi	X̄±SS	7,01±1,21	6,82±1,41	6,95±1,27	0,177[b]
	Ortanca (Minimum-maksimum)	7,00 (4,00-11,00)	7,00 (4,00-10,00)	7,00 (4,00-11,00)	
Günlük derin uyku süresi	X̄±SS	5,49±1,83	5,51±1,95	5,50±1,87	0,725[b]
	Ortanca (Minimum-maksimum)	6,00 (0,00-11,00)	6,00 (2,00-10,00)	6,00 (0,00-11,00)	
Düzenli ilaç alımı, n (%)	Yok	563 (91,8)	50 (8,2)	613 (64,7)	**<0,001**[a]
	Var	120 (35,8)	215 (64,2)	335 (35,3)	
Abur cubur tüketimi, n (%)	Ara sıra	469 (70,1)	200 (29,9)	669 (70,6)	0,131[a]
	Sıklıkla	141 (77,0)	42 (23,0)	183 (19,3)	
	Çok sık	37 (71,2)	15 (28,8)	52 (5,5)	
	Her zaman	36 (81,8)	8 (18,2)	44 (4,6)	
Stres, n (%)	Yok	97 (71,9)	38 (28,1)	135 (14,2)	**<0,001**[a]
	Bazen	449 (79,8)	114 (20,2)	563 (59,4)	
	Sıklıkla	106 (65,0)	57 (35,0)	163 (17,2)	
	Her zaman	31 (35,6)	56 (64,4)	87 (9,2)	
Kan basıncı seviyesi, n (%)	Düşük	28 (100,0)	0 (0,0)	28 (3,0)	**<0,001**[a]
	Normal	578 (82,0)	127 (18,0)	705 (74,4)	
	Yüksek	77 (35,8)	138 (64,2)	215 (22,6)	
İdrara çıkma sıklığı, n (%)	Normal	496 (75,0)	165 (25,0)	661 (69,7)	**0,002**[a]
	Oldukça fazla	187 (65,2)	100 (34,8)	287 (30,3)	

[a]Pearson ki-kare testi; [b]Mann-Whitney U testi; *Var/yok sütunlarında satır yüzdesi, toplam sütununda sütunda yüzdesi verilmiştir; SS: Standart sapma; BKİ: Beden kitle indeksi.

Şekil 2'de gain ratio değişken önemi testi kullanılarak genel model (a), birinci küme (b) ve ikinci küme (c) için değişken önemine, değişkenlerin diyabete etkisine bakıldı. Tüm veri için değişken önemi testi sonucu ve klinik öneme göre değerlendirildiğinde model düzenli ilaç alımı, kan basıncı seviyesi, yaş, günlük uyku süresi, ailede diyabet öyküsü ve stres değişkenlerinden oluştu. Sonuç olarak genel model için çalışmaya 7 değişken (6 bağımsız, 1 bağımlı değişken) dâhil edildi ve makine öğrenmesi analizleri bu değişkenler kullanılarak yapıldı (Şekil 2a).

Birinci küme için değişken önemi testi sonucu ve klinik öneme göre değerlendirildiğinde model düzenli ilaç alımı, günlük uyku süresi, yaş, kan basıncı seviyesi, stres, günlük derin uyku süresi ve idrara çıkma sıklığı değişkenlerinden oluştu. Sonuç olarak genel model için çalışmaya 8 değişken (7 bağımsız, 1 bağımlı değişken) dâhil edildi ve makine öğrenmesi analizleri bu değişkenler kullanılarak yapıldı (Şekil 2b).

İkinci küme için değişken önemi testi sonucu ve klinik öneme göre değerlendirildiğinde model düzenli ilaç alımı, ailede diyabet öyküsü, kan basıncı seviyesi, yaş ve günlük derin uyku süresi değişkenlerinden oluştu. Sonuç olarak genel model için çalışmaya 6 değişken (5 bağımsız, 1 bağımlı değişken) dâhil edildi ve makine öğrenmesi analizleri bu değişkenler kullanılarak yapıldı (Şekil 2c).



ŞEKİL 2: Gain ratio yöntemine göre değişken önemi sonuçları.
BKİ: Beden kitle indeksi

Birinci küme genel model ile kıyaslandığında günlük uyku süresi ve stres değişkenleri daha önemli hâle geldi ve genel modelde yer almayan günlük derin uyku süresi ve idrara çıkma sıklığı değişkenleri bu modelde yer aldı. İkinci küme genel model ile kıyaslandığında ailede diyabet öyküsü değişkeni daha önemli hâle geldi ve genel modelde yer almayan günlük derin uyku süresi değişkeni model dâhil edildi.

Tahmin performanslarını değerlendirmek için lojistik regresyon, çok katmanlı algılayıcı, karar tablosu, karar destek makinesi, rastgele orman, J48, IBk, torbalama ve oylanmış algılayıcı yöntemleri kullanıldı. Bu yöntemler arasından en iyi sonuca genel model ve ikinci küme için rastgele orman yöntemi ile birinci küme için ise J48 yöntemi ile ulaşıldı (Tablo 2).

Genel model için DSO %93,4, birinci küme için %93,3, ikinci küme için ise %93,2 olarak bulundu. Genel model ile karşılaştırıldığında, iki küme sonuçları ile benzer olduğu gözlendi. Modeller ile Tip 2 diyabet olacak ve olmayacak hastalar için tahmin sonuçlarına bakıldığında, birinci kümede Tip 2 diyabet görülecek tahmini yapılan hastalara ait DSO'da genel modele göre artış gözlendi. Genel modele ait DSO %83,0 iken bu oran birinci kümeye ait model ile %90,5 olarak bulundu. İkinci kümede Tip 2 diyabet görülmeyecek tahmini yapılan hastalara ait DSO'da da genel modele göre artış gözlendi. Genel modele ait DSO %97,4 iken bu oran ikinci kümeye ait model ile %97,6 olarak bulundu (Tablo 2).

TABLO 2: Modeller için performans ölçütleri.

Veri setleri	Kategoriler	Performans ölçütleri				
		DSO	F-ölçütü	MCC	PRC alanı	AUC
Orjinal veri seti	Tip 2 DM yok	0,974	0,955	0,832	0,981	0,966
	Tip 2 DM var	0,830	0,875		0,939	
	Genel	0,934	0,932		0,969	
Küme 1	Tip 2 DM yok	0,946	0,950	0,847	0,953	0,938
	Tip 2 DM var	0,905	0,897		0,907	
	Genel	0,933	0,933		0,938	
Küme 2	Tip 2 DM yok	0,976	0,957	0,800	0,973	0,946
	Tip 2 DM var	0,784	0,840		0,910	
	Genel	0,932	0,930		0,959	

DM: Diabetes mellitus; DSO: Doğru sınıflama oranı; MCC: Matthews korelasyon katsayısı; PRC: Precision recall curve; AUC: Eğrinin altındaki alan.

TARTIŞMA

İdeal küme sayısı bulunup veri kümelere bölündüğünde, genel olarak kümelerin her birine bakıldığında klinik olarak beraber önemli olan değişkenler bir araya geldi, bu sayede analizler daha spesifik grubu temsil edecek şekilde yapıldı. Aynı zamanda, kümelere ait örneklem sayısı genel modelin neredeyse yarısı kadar olsa da sonuçlar değerlendirildiğinde, hipotezimizle paralel olarak kümeler için oluşturulan modeller daha iyi sınıflama gücüne sahipti.

Literatürde tıp alanında kümeleme yöntemleri kullanarak hastaları alt gruplara ayıran ve bu gruplarda sınıflama yöntemi kullanarak, tahminleme performanslarını gösteren çalışma sayısı sınırlıdır. Yapılan çalışmalarda genellikle hastalar özelliklerine göre alt gruplara ayrılmış ve bu gruplarda karşılaştırmalar yapılarak bulunan farklılıklar sunulmuştur.

Tao ve ark. Tip 2 diyabetli 908 hasta üzerinde yaptıkları çalışmada, hastaları kümeleme yöntemi ile özelliklerine göre 4 farklı alt gruba ayırdı ve alt grupların istatistiksel ve klinik farklılıklarla yeterince ayırt edilebildiğini gösterdi.[9] Ghassib ve ark. 6 ve daha fazla dişe sahip 1.222 katılımcı üzerinde K-ortalama kümele yöntemi kullanarak yaptıkları analiz sonucunda, hastaları özelliklerine göre 5 kümeye ayırdı ve verilerinin kümelenmesinin, periodontitis durumuyla istatistiksel olarak anlamlı şekilde ilişkili olan özellikleri etkili bir şekilde tanımladığını ve bu sayede maliyetli klinik muayeneler olmaksızın periodontitis için yüksek risk taşıyan alt popülasyonları tespit ettiğini bildirdi.[10] Ames ve ark. erişkin omurga deformitesi olan 575 hasta üzerinde yaptıkları çalışmada, denetimsiz hiyerarşik kümeleme yöntemi ile hastaları özelliklerine göre koronal düzlem deformitesi olan genç hastalar (n=195), daha önce omurga ameliyatı geçirmiş daha yaşlı hastalar (n=157) ve daha önce omurga ameliyatı geçirmemiş daha yaşlı hastalar (n=218) olmak üzere 3 alt gruba ayırdı. Alt gruplarda yapılan analizler sonucunda, kümeleme yönteminin ameliyat öncesi karar vermeyi artırabilecek veri modellerini belirlemede fayda sağlayacağı ve oluşturulacak yapay zekâ tabanlı sınıflandırma ile cerrahları hangi tedavi modellerinin en düşük riskle en iyi iyileşmeyi sağladığı konusunda eğiterek tedavi optimizasyonunu kolaylaştırabileceği sonucu vardı.[11] Stoitsas ve ark. travma tedavisi için hastaneye başvuran 4.883 kişiyi dâhil ettikleri çalışmada, 3 farklı kümeleme yöntemi kullanarak hastaları farklı sayıda alt gruplara ayırmış ve alt grupların tahmin performanslarını farklı sınıflandırma yöntemleri kullanarak incelemiştir. En yüksek DSO'ya (%91,3) ise Deepgmm yöntemi ile ulaşmışlardır.[12] Bu çalışmada, veri seti 2 alt gruba ayrılmış ve tahminlemeler alt gruplarda ve tüm hasta içeren veri setinde ayrı ayrı yapılmıştır. Literatürle uyumlu olarak, kümeleme sonucunda elde edilen iki kümenin hastalara ait belirgin özellikleri iyi temsil ettiği görülmüştür ve kümelere ait tahminlemenin genel modelden daha iyi olduğu bulunmuştur.

SONUÇ

Hastalara ait yaş, cinsiyet, ek hastalık gibi demografik özellikler ve çeşitli çevresel faktörler hasta verilerinde heterojenliğe yol açmaktadır. Hasta verilerini, benzer özelliklere sahip kümelere ayırmadan yapılan analizler ise risk faktörlerinin eksik/yanlış belirlenmesine ve oluşturulan modelin yanlış olmasına neden olmaktadır. Mevcut çalışmada, bu durumlar göz önünde bulundurularak yapılan kümeleme analizi sonrasında hastalara ait spesifik özelliklerin daha iyi belirlendiği ve küme bazlı sınıflandırmayla da oluşturulan modellerinin daha iyi sınıflama performansına sahip olduğu gözlenmiştir.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Bu çalışma tamamen yazarın kendi eseri olup başka hiçbir yazar katkısı alınmamıştır.

KAYNAKLAR

1. Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Stat Med. 2005;24(9):1307-19. PMID: 15685717.
2. Ramaekers BL, Joore MA, Grutters JP. How should we deal with patient heterogeneity in economic evaluation: a systematic review of national pharmacoeconomic guidelines. Value Health. 2013;16(5):855-62. PMID: 23947981.
3. Ali M, Salehnejad R, Mansur M. Hospital heterogeneity: what drives the quality of health care. Eur J Health Econ. 2018;19(3):385-408. PMID: 28439750; PMCID: PMC5978923.
4. Hernández A. Heterogeneity of patients in clinical trials: Subgroup analysis and covariate adjustment in cardiovascular and neurosurgical trials [Master thesis]. Rotterdam: Erasmus University Rotterdam; 2006. [Cited: 2023]. Available from: https://repub.eur.nl/pub/7456/
5. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355(9209):1064-9. PMID: 10744093.
6. Lader EW, Cannon CP, Ohman EM, Newby LK, Sulmasy DP, Barst RJ, et al; American College of Cardiology Foundation. The clinician as investigator: participating in clinical trials in the practice setting. Circulation. 2004;109(21):2672-9. PMID: 15173050.
7. Li X, Jiao H, Li D. Intelligent medical heterogeneous big data set balanced clustering using deep learning. Pattern Recognit. Lett. 2020;138(1):548-55. https://www.sciencedirect.com/science/article/abs/pii/S0167865520303329?via%3Dihub
8. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput. Sci. 2020;167:706-16. https://www.sciencedirect.com/science/article/pii/S1877050920308024
9. Tao R, Yu X, Lu J, Shen Y, Lu W, Zhu W, et al. Multilevel clustering approach driven by continuous glucose monitoring data for further classification of type 2 diabetes. BMJ Open Diabetes Res Care. 2021;9(1):e001869. PMID: 33627315; PMCID: PMC7908294.
10. Ghassib IH, Batarseh FA, Wang HL, Borgnakke WS. Clustering by periodontitis-associated factors: A novel application to NHANES data. J Periodontol. 2021;92(8):1136-50. PMID: 33315260.
11. Ames CP, Smith JS, Pellisé F, Kelly M, Alanay A, Acaroğlu E, et al; European Spine Study Group, International Spine Study Group. Artificial intelligence based hierarchical clustering of patient types and intervention categories in adult spinal deformity surgery: towards a new classification scheme that predicts quality and value. Spine (Phila Pa 1976). 2019;44(13):915-26. PMID: 31205167.
12. Stoitsas K, Bahulikar S, de Munter L, de Jongh MAC, Jansen MAC, Jung MM, et al. Clustering of trauma patients based on longitudinal data and the application of machine learning to predict recovery. Sci Rep. 2022;12(1):16990. PMID: 36216874; PMCID: PMC9550811.